1-1-2021

Predicting self-intercepted medication ordering errors using machine learning

Christopher Ryan King

Joanna Abraham

Bradley A. Fritz

Zhicheng Cui

William Galanter

See next page for additional authors

Authors

Christopher Ryan King, Joanna Abraham, Bradley A. Fritz, Zhicheng Cui, William Galanter, Yixin Chen, and Thomas Kannampallil

PLOS ONE

RESEARCH ARTICLE

Predicting self-intercepted medication ordering errors using machine learning

Christopher Ryan King[1], Joanna Abraham[1,2], Bradley A. Fritz[1], Zhicheng Cui[3], William Galanter[4,5], Yixin Chen[3], Thomas Kannampallil[1,2]*

1 Department of Anesthesiology, Washington University School of Medicine, Saint Louis, Missouri, United States of America, **2** Institute for Informatics, Washington University School of Medicine, Saint Louis, Missouri, United States of America, **3** Department of Computer Science, McKelvey School of Engineering, Washington University in St Louis, Saint Louis, Missouri, United States of America, **4** Department of Medicine, College of Medicine, University of Illinois at Chicago, Chicago, Illinois, United States of America, **5** Department of Pharmacy Systems, Outcomes and Policy, College of Pharmacy, University of Illinois at Chicago, Chicago, Illinois, United States of America

* thomas.k@wustl.edu

Citation: King CR, Abraham J, Fritz BA, Cui Z, Galanter W, Chen Y, et al. (2021) Predicting self-intercepted medication ordering errors using machine learning. PLoS ONE 16(7): e0254358. https://doi.org/10.1371/journal.pone.0254358

Editor: Usman Qamar, National University of Sciences and Technology (NUST), PAKISTAN

Received: October 27, 2020

Accepted: June 27, 2021

Published: July 14, 2021

Data Availability Statement: This study is protected by the institutional review protocols approved by the University of Illinois (IRB #2015-0767) and from the human research protection office (HRPO) of Washington University (IRB #2018-06061). The data used for this study contains protected health information and cannot be shared without explicit data use agreements in place. Data requests can be sent to either of the following institutional offices: Institutional Review Board at Washington University, Human Research Protection Office 660 South Euclid Avenue, Campus Box 8089, St. Louis, MO 63110,

Abstract

Current approaches to understanding medication ordering errors rely on relatively small manually captured error samples. These approaches are resource-intensive, do not scale for computerized provider order entry (CPOE) systems, and are likely to miss important risk factors associated with medication ordering errors. Previously, we described a dataset of CPOE-based medication voiding accompanied by univariable and multivariable regression analyses. However, these traditional techniques require expert guidance and may perform poorly compared to newer approaches. In this paper, we update that analysis using machine learning (ML) models to predict erroneous medication orders and identify its contributing factors. We retrieved patient demographics (race/ethnicity, sex, age), clinician characteristics, type of medication order (inpatient, prescription, home medication by history), and order content. We compared logistic regression, random forest, boosted decision trees, and artificial neural network models. Model performance was evaluated using area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). The dataset included 5,804,192 medication orders, of which 28,695 (0.5%) were voided. ML correctly classified voids at reasonable accuracy; with a positive predictive value of 10%, ~20% of errors were included. Gradient boosted decision trees achieved the highest AUROC (0.7968) and AUPRC (0.0647) among all models. Logistic regression had the poorest performance. Models identified predictive factors with high face validity (e.g., student orders), and a decision tree revealed interacting contexts with high rates of errors not identified by previous regression models. Prediction models using order-entry information offers promise for error surveillance, patient safety improvements, and targeted clinical review. The improved performance of models with complex interactions points to the importance of contextual medication ordering information for understanding contributors to medication errors.

hrpo@wustl.edu 1-800-438-0445 Institutional Review Board at University of Illinois, Office of Protection of Research Subjects, 1737 West Polk Street, Suite 310, MC 672, Chicago, IL 60612, uicirb@uic.edu 1-312-996-1711.

Funding: This project was supported in part by grant (R21HS025443) from the Agency for Healthcare Research and Quality to JA. The content is solely the responsibility of the authors and does not represent the official views of the Agency for Healthcare Research and Quality [https://www.ahrq.gov/]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Computerized provider order entry (CPOE) systems streamline medication ordering process by creating standardized templates for the entry of legible, accurate, and complete medication orders, thereby mitigating the potential for medication errors [1–6]. By integration with electronic health record (EHR) systems, CPOE systems promote the coordination of medication ordering [7], allow real-time collaboration among care team members for medication administration, delivery and monitoring [8], and reduce the potential for misinterpretation of orders [9]. Tight coupling of the EHR with an ordering system has also led to the development of clinical decision support tools for alerting clinicians regarding improperly composed orders, clinically inappropriate orders, duplicate orders [10], wrong patient orders, and formulary non-compliance [11].

With the widespread use of CPOE systems, the volume of medications orders, and correspondingly, medication errors have increased exponentially. Rough estimates suggest that 25–30 orders are placed per inpatient admission, with nearly 4–6 additional orders per patient per day [12, 13]. Clinician interactions with CPOE systems are a source of medication errors [14–17]; errors during CPOE use account for 6–25% of detected medication errors in hospitalized patients [15]. In spite of its prevalence, much of the prior research on the causes of medication ordering errors has relied on small samples from retrospective chart reviews, clinician self-reports, analysis of malpractice claims data, and survey-based studies [18, 19]. One of the larger analysis of CPOE-based errors used manual reviews to categorize over 10,000 reported errors drawn from a national database [20], classifying the causes of such errors. Although such analyses are useful in understanding the sources of CPOE-based medication errors, these databases include limited details regarding the context of a reported error with considerable variability regarding the content of reported errors. Additionally, because of the lack of matched control (non-error) orders, they cannot be used for developing prediction models.

Previous estimates of medication errors have relied on self-reports, which often capture only the most severe errors or errors that lead to patient harm. Studies using pharmacy-based reviews are also likely to undercount errors, as orders that seem "plausible" go undetected even if they are on an incorrect patient. One approach that has received traction is the use of CPOE-integrated tools that help clinicians record self-intercepted errors within their workflow (e.g., [14, 21, 22]). Such intercepted ordering errors, although still a fraction of overall medication errors, provide considerable advantages: first, the volume of self-intercepted errors within a CPOE system is exponentially larger than manually reported errors to external incident reporting systems. Second, given that self-intercepted errors are recorded within the CPOE workflow, additional information regarding the context of the order is also easily available.

We previously described a dataset of self-intercepted errors and identified associated factors using bivariate and regression analyses [22]. That analysis identified several highly plausible risk factors for error (or interception) such as order type (e.g., inpatient, prescription) and prescriber role (e.g., physician, pharmacist, student, nurse). However, exploratory data analysis and manual stepwise modeling with massive scale data, rare outcomes, and large numbers of variables is inherently limited in its scope.

In this paper, we propose the use of machine learning (ML) approaches for characterizing the risk factors associated with medication ordering errors. Towards this end, we evaluated the performance of multiple ML methods on a large dataset of self-intercepted medication ordering errors. Such automated predictions afford opportunities for characterizing the potential causes and sources medication errors as well as allowing for automated error surveillance. We also discuss methodological advantages of using ML for predicting medication ordering errors

and opportunities for using these approaches to guide patient safety efforts that are targeted towards medication orders within high-risk contexts.

Method

A TRIPOD checklist was used for the development of the prediction model and is included in the S1 Checklist.

Setting

Medication orders generated over a 6-year period (2006–2011) at the University of Illinois Hospital and Health Sciences System (UI Health) were studied. UI Health is a 495-bed tertiary, urban academic medical center. Orders were placed using Cerner Powerchart and Firstnet. Pharmacists or nurses entered orders based on verbal, written, or protocol orders from physicians; students entered orders requiring physician approval. Additional details of the data collection are found in a previous report [22]. This study was approved by the Institutional Review Board of Washington University in St Louis and University of Illinois at Chicago with a waiver of consent.

Medication order voiding is a CPOE-integrated function for physicians, pharmacists, nurses, and students to *self-intercept* and remove erroneous medication orders. As opposed to Medication Error Reporting Systems (MERS) reports, voiding can be performed within the ordering workflow, allowing clinicians to document errors without accessing external systems or requiring providing detailed descriptions of the error. Previous studies have shown that voided orders are a good proxy for medication errors, with voided medication orders having a 70±10% positive predictive value of being an error [14, 22]. In the paper, medication ordering errors refer to the errors identified using the voiding process. Although a field for choosing a reason for voiding existed, we have found it to be unreliable and did not use it for this analysis [14].

Data

The outcome variable or label was order status (i.e., whether an order was voided or not). Other variables (predictors or features) included: patient demographics (race/ethnicity, sex, age), clinician type, type of medication order (inpatient, prescription, home medication by history), order date and time, and order content. Reported race was categorized into: White, Black, Hispanic, and other. Order type was classified as: inpatient (i.e., medication order for a hospitalized patient), prescriptions, and home medications by history (a non-actionable record of a medication that a patient was taking at home and was not recorded as a prescription). Time was categorized as: day (7AM–5PM), night (5PM–12AM), or overnight (12AM–7AM); Weekday was classified as: work weekday (Monday through Friday) or weekend (Saturday or Sunday). Clinician type was categorized as: physician, pharmacist, nurse, student, or other. Order content included: drug name, route, strength, volume, frequency, and dispensing unit (e.g., tablet, cap box). For each unique medication and route combination, doses were z-scored. All data were retrieved from the EHR using custom queries.

Statistical analysis

We constructed the following models: logistic regression (LR), decision tree (DT), random forest (RF), gradient boosted decision tree (GBDT), deep embedding logistic regression (DELR), and multilayer perceptron (MLP) models. DELR is an extension of logistic regression in which each feature is fed into its own deep, narrow neural network to allow nonlinear transformation

prior to entry into the logistic regression model [23]. LR, DT, RF, GBDT, and DELR were selected due to the ease of determining feature importance in each of these model architectures. Despite its generally poor interpretability, MLP was selected for its high accuracy of neural networks other applications such as image classification [24] and object detection [25].

The dataset was split into training, validation, and testing sets with a ratio of roughly 7:1:2. Because of the large sample size, repeated k-fold splits were not constructed. Missing categorical variables were added as a level of the feature. For models (such as logistic regression and MLP) that do not natively handle missing quantitative predictors, missing values were imputed with the mean observed value and a "missing value" indicator feature was concatenated. For each model, hyperparameters were searched over a grid of plausible values with tests for expansion at each endpoint, where possible. Hyperparameters that resulted in the highest area under the receiver operating characteristic curve (AUROC) in the validation set were selected and carried forward to the test set evaluation.

As the dataset was highly imbalanced (0.5% of medication orders were voided), we utilized two techniques to address class imbalance in the training set [26]. First, class weights were set as inversely proportional to their proportion. Second, we up-sampled positive training cases by 201 times, equalizing proportions. For each model architecture, both of these techniques were applied to the training set. Neither weighting nor up-sampling were performed in the validation or testing sets. The choice of class imbalance resolution technique with better performance in the validation set was used in the training model to be evaluated on the test set. That is, we treated reweighting versus up-sampling as a hyperparameter.

LR was tested with or without L2 (ridge) and L1 ratio (lasso) penalty selected from 1000, 100, 10, 1,.1 and 0,.1,.25,.5,.75, 1. LR with pairwise interactions and filtering by L1 penalty was attempted, but this classifier was difficult to optimize during training, tested very poorly, and is not reported. For DT, tree depth ranged from 3 to 8, and minimum sample split was selected from 2, 1000, 2000, 10,000, and 20,000 (a minimum sample split of 2 represents no early termination). For RF and GBDT, the number of decision trees was 100, 200, 300, 400, or 500, decision tree depth ranged from 2 to 8, and minimum sample split ranged from 2 to 20,000. For DELR, each transformation network had depth of 2 to 6 layers and width of 4 to 6 nodes. For MLP, network depth ranged from 2 to 6 layers and hidden layer width was selected from 32, 64, or 128 neurons. For both DELR and MLP, two optimizers (stochastic gradient descent with learning rate 0.1 or 0.01 and Adam optimizer with learning rate 0.001 [24]) were tested. Presence or absence of batch normalization was also tested.

Once the hyperparameters were fixed, model performance in the testing set was quantified using the AUROC metric. In highly imbalanced datasets, AUROC can sometimes be deceptive on evaluating the model performance [25]. As such, we also constructed the precision-recall curve for each model and calculated the area under the precision-recall curve (AUPRC), which is a measure of average precision. Confidence intervals for AUROC and AUPRC were created using the non-parametric logit method [27] in MatlabAUC package [28]. Each classifier was re-calibrated using isotonic regression in the validation cohort [29].

For each model, the most important features contributing to the predictions were identified. For LR, feature importance was defined by the absolute value of the regression coefficient. For DELR, feature importance was defined as the product of feature embedding value and regression coefficient. In models with deep interactions, "explainability" was approached in multiple ways. Global feature importance was assessed, for example, by permutation; however, in complex models the direction and magnitude of effect for a feature in a specific example depends on the context. For DT, feature importance was defined as the weighted entropy decrease on that feature during training phase. For RF and GBDT, feature importance was determined for the component DTs and averaged across all DTs in the model. For MLP,

feature importance was quantified using backpropagation-based salience detection (integrated gradients) [30]. In this technique, the gradient of the output prediction with respect to the input feature values for each case was calculated, and the gradient with respect to each feature was averaged across the population. Integrated gradients and Shapely values were calculated for each observation, permitting local measures of feature relevance [31]. To highlight important deep interactions, we used a large global decision tree that can approximate GBDT and RF fitted surfaces [32] when appropriately supervised. We also permuted features to assess global importance to model fit. All analyses were conducted using Python version 3.7, unless otherwise specified.

Results

The dataset included 5,804,192 orders, of which 28,695 (0.5%) were voided orders. Characteristics of the orders are shown in Table 1. Nearly two-thirds of orders were inpatient medication orders, and most orders were placed during day shifts. Overall, there were more order voiding on orders created by students (4%) or by nurses (i.e., as verbal orders, 1%).

Comparing model performance

Hyperparameters that resulted in superior AUROC in the validation set were identified. For LR, presence of L2 penalty was selected. For DT, tree depth of 8 and minimum sample split of 20,000 were selected. For RF, decision tree number of 500, decision tree depth of 8, and minimum sample split of 20000 were selected. For GBDT, decision tree number of 100, decision tree depth of 8, and minimum sample split of 2000 were selected. For DELR, transformation network depth of 4 layers, hidden layer width of 5 neurons were selected and batch normalization was enabled. For MLP, network depth of 3 layers and hidden layer width of 32 neurons were selected. For DELR and MLP, stochastic gradient descent with learning rate of 0.1 and 0.01 was selected respectively. The inverse weighting method for addressing class imbalance performed better than the up-sampling method for the DT and GBDT, while the up-sampling method performed better for the LR, RF, DELR and MLP.

After fixing the hyperparameters, model performance in the testing set was computed (see Table 2). GBDT achieved the highest AUROC and AUPRC across all models. RF had approximately the same accuracy as DT, with marginally better AUPRC and marginally worse AUROC. LR and DELR demonstrated the poorest performance of all the models, especially in the AUPRC metric. Receiver operating characteristic and precision-recall curves for all models are shown in Fig 1. All models had acceptable calibration in the test set; the calibration curve for each model is shown in the S1 Checklist.

Model interpretation

We also investigated the potential for deriving meaningful explanations from the developed models. Fig 2 represents a partial decision tree for voided orders with an overall depth of 7, 54 decision nodes, and 63 leaf nodes. Displayed leaf nodes had >100 examples in the testing data; error rates within this tree varied between 0.003% to 46%. A large segment of the presented tree includes medication ordering without a value for the feature "volume dose unit" (e.g., milliliters, tablets, capsules, allowed to differ from "dispense unit" and "strength unit"). For example, the top, left-most leaf node represents orders related to the medication, nalbupine, with a volume dose unit missing or unspecified. Such orders have a high rate of being voided (~34%). This is potentially because these orders are often added outside of an order set (e.g., for pain management after surgery) or using a prespecified "order sentence." The second node

Table 1. Characteristics of orders in the dataset.

Variable	Non-Voided Orders (N = 5,775,497)	Voided Orders (N = 28,695)
Age (mean, SD)	45.5 (22.3)	46.5 (22.2)
Age Group		
0–19	793,672 (14%)	3,861 (13%)
20–39	1,333,142 (23%)	6,131 (20%)
40–49	826,481 (14%)	3,741 (12%)
50–59	1,127,423 (20%)	7,941 (26%)
60–69	899,144 (16%)	4,888 (16%)
70–79	540,829 (9%)	2,817 (9%)
>80	254,806 (4%)	1,296 (4%)
Sex		
Female	3,444,161 (60%)	16,910 (59%)
Male	2,330,629 (40%)	11,778 (41%)
Race/Ethnicity		
Black	2,938,539 (51%)	14,075 (49%)
Caucasian	1,217,607 (21%)	6,794 (24%)
Hispanic	729,051 (13%)	3,550 (12%)
Other	890,300 (15%)	4,276 (15%)
Order Type		
Normal Order	3,620,733 (63%)	17,455 (61%)
Prescription/Discharge Order	1,689,052 (29%)	5,056 (18%)
Recorded/Home Meds	465,708 (8%)	6,184 (21%)
Shift		
Day	3,957,683 (69%)	19,991 (70%)
Night	1,177,477 (20%)	6,148 (21%)
Overnight	640,336 (11%)	2,556 (9%)
Title of Ordering Provider		
Nurse	546,650 (9%)	5,684 (20%)
Pharmacist	627,963 (11%)	2,382 (8%)
Physician	3,954,439 (68%)	15,769 (55%)
Student	38,050 (1%)	1,655 (6%)
Other	608,395 (11%)	3,205 (11%)
Normalized Drug Dose		
Normalized Strength (mean, SD)	2.6e-4 (0.99)	-5.4e-2 (1.16)
Normalized Volume (mean, SD)	5.7e-6 (0.98)	-1.9e-3 (2.66)

SD = standard deviation

https://doi.org/10.1371/journal.pone.0254358.t001

highlights orders that are entered by a medical student; for inpatient orders, 24% of such orders were likely voided.

The root node, missing volume dose unit, also highlights a contextual aspect related to the source of the medication errors. Medication orders were sometimes created using "order sentences," where medication-related information was standardized and incorporated into an order. For instance, for the common blood pressure medication, amlodipine, an order sentence is available for a standard "10 mg, PO ("per os" or by mouth), daily." If the clinician wanted an unusual dose or frequency, they would need to enter it as text; for example, "2.5 mg, PO, 2 AM and 1 PM." Another situation where this occurs is when no order sentences are

Table 2. Model performance using the test data set (N = 1160839).

Model	AUROC	AUPRC	NORMAL AUROC	NORMAL AUPRC
	[95% CI]	[95% CI]	[95% CI]	[95% CI]
LR	0.7468	0.0225	0.7518	0.0223
	[0.7398, 0.7536]	[0.0165, 0.0306]	[0.7430, 0.7604]	[0.0160, 0.0338]
DT	0.7769	0.0407	0.7810	0.0416
	[0.7590, 0.7938]	[0.0265, 0.0662]	[0.7581, 0.8023]	[0.0243, 0.0702]
RF	0.7635	0.0242	0.7679	0.0248
	[0.7568, 0.7701]	[0.0183, 0.0318]	[0.7594, 0.7762]	[0.0177, 0.0347]
DELR	0.7561	0.0230	0.7617	0.0239
	[0.7493, 0.7627]	[0.0172, 0.0308]	[0.7532, 0.7700]	[0.0168, 0.0339]
MLP	0.7868	0.0562	0.7888	0.0567
	[0.7803, 0.7932]	[0.0501, 0.0630]	[0.7804, 0.7969]	[0.0490, 0.0655]
GBDT	0.7968	0.0647	0.8005	0.0661
	[0.7905, 0.8030]	[0.0587, 0.0713]	[0.7925, 0.8083]	[0.0586, 0.0745]

https://doi.org/10.1371/journal.pone.0254358.t002

available, usually for rarely used medications or medications with highly varying doses. During such situations, the volume dose unit is unlikely to be defined, potentially leading to voided orders. We found that such orders with missing volume dose unit have a higher likelihood of being errors as highlighted by the sample decision tree (Fig 2).

Discussion

To the best of our knowledge, this is the first study to utilize a large, routinely collected data set of self-intercepted medication ordering errors to forecast error status and identify associated features of these errors. Using a database of >5 million medication orders, we compared five classification algorithms for predicting medication ordering errors based primarily on the contextual features associated with a medication order. We found that GBDT was the top performing model (AUROC = 0.797) and was able to predict errors with a 10% PPV at a sensitivity of 20%. Based on a simplified classification rule using the DT algorithm, most of the voided orders had a low risk (rate ~0.3%), and about 5% with a much higher risk (rate >4%). The models highlighted the association of errors to known risk-increasing features of an order

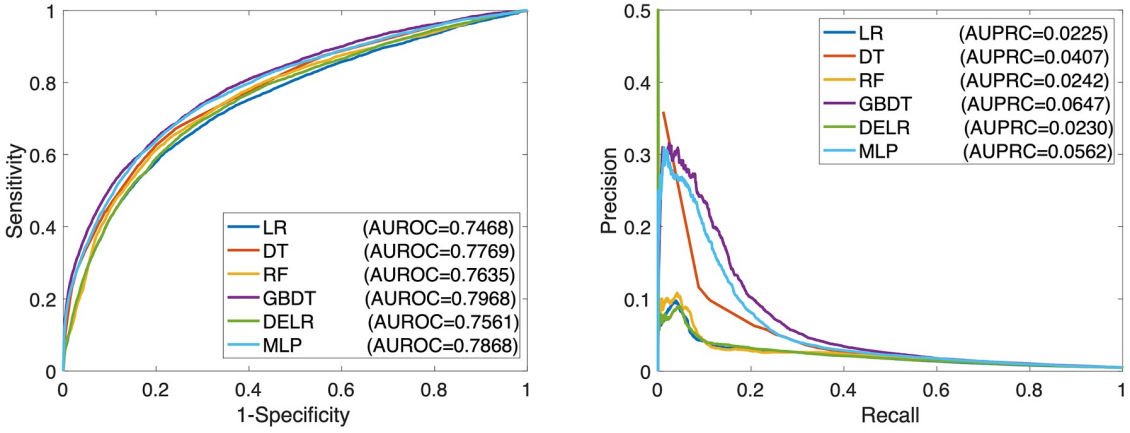

Fig 1. Model performance. Receiver operating characteristic curves (left panel) and precision-recall curves (right panel) demonstrating performance of each model in the testing set.

https://doi.org/10.1371/journal.pone.0254358.g001

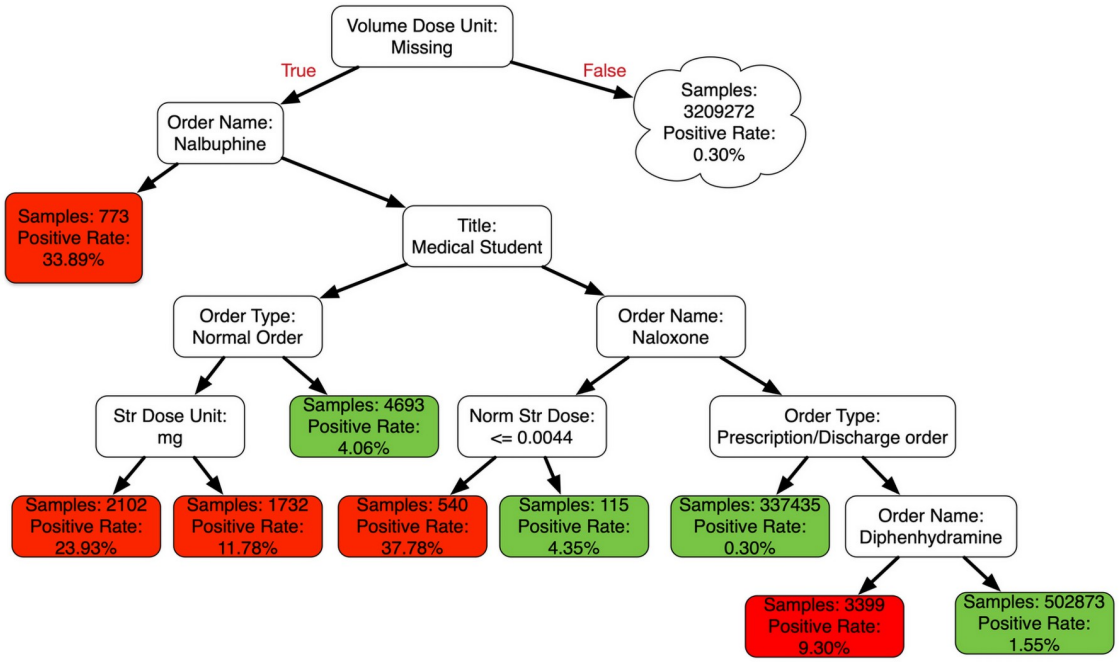

Fig 2. Example partial tree from the DT model showing the causal patterns for medication errors. In this tree, if the decision is false (i.e., not a voided order), take the right branch; similarly, if the decision is true (i.e., voided order), take the left branch. Red and green boxes highlight the positive rates of greater than and less than 5% respectively. Note that this is a small portion of the total tree.

https://doi.org/10.1371/journal.pone.0254358.g002

such as student-formulated orders, and to previously unknown and likely local setting-based features such as missing fields (e.g., missing volume dose unit).

The methodological approach of using ML algorithms for predicting medication errors has two potential applications. First, it is possible to identify factors associated with order entry errors that potentially represent generalizable knowledge for mitigating such errors. In Fig 2, orders that had a volume dose unit specified, most likely arising from standardized order sentences, had a lower likelihood of being voided. It is not surprising that orders that do not have a pre-defined dose unit would be associated with more errors. Such orders require more "clicks" as well as clinician recall or the need to look up dose units, fragmenting the ordering workflow. Further investigation of the use of order sentences is a promising line of investigation. Similarly, the high rate of voided student and verbal orders likely represent real and generalizable finding. Again, although not surprising, these findings have implications for training of future users. Unlike raw counts, model-based outputs are adjusted for related features, reducing the effects of confounding.

Second, ML-based medication ordering error identification offers opportunities to guide patient safety efforts that are targeted towards medication orders within high-risk contexts. "Context" has two potential meanings here. First, is the modeling interpretation of "interactions between features." Models that did not include interactions, such as logistic regression, were less accurate, suggesting that multiple features forming a "context" is potentially required to identify medication ordering errors. In our example (Fig 2), low naloxone doses in particular were likely to be voided, suggesting that these are a confusing or easily mistaken option rather than naloxone itself as a problematic drug.

The second meaning of "context" is that many of these findings are likely situated within the context of the customized CPOE of our dataset and would neither be apparent or relevant to a multi-institutional or national database. To revisit the above example of missing volume

dose unit, other CPOEs may not use the same structure to indicate "written without an order sentence/order set." More specifically, the high rate of voided nalbuphine orders likely reflects ordering options that are not working as intended. In the limiting case, it is also possible to narrow the explanation to an individual clinician's medication ordering characteristics [33]. Such an approach can account for the context of historical errors that are aligned with a specific clinician's ordering characteristics or characteristics of a medication (e.g., a high-risk drug). These local findings are important to find issues in a specific CPOE implementation and user behaviors that can affect patient safety. Our methodological approach, hence, can be applied directly to hospital-scale data.

Finally, there may be value in using model-based predictions to target orders for further scrutiny. With the use of voiding function (or similar self-interception detection techniques), observed counts of errors can be a valuable source for patient safety investigations. However, in any self-report mechanism false negatives are also likely. Orders sharing characteristics with frequently voided ones (i.e., high forecasted probability) could be prospectively reviewed even if the yield is below the threshold to justify clinician-facing decision support. In our case, at a 20% sensitivity, the 10% positive predictive value implies a 20-fold decrease in the number of orders to review to discover one error. Drawing these enriched samples is a prerequisite for feasible patient safety review given that the total error rate is very low, especially for smaller clinical units within a hospital. Additionally, if units within a hospital have differential accuracy in error reporting, then forecasted error rates may be necessary to identify safety issues. Another way to understand this application is that the predicted error rate can act as a prior for a more stable estimate than the naïve reported rate in a small or noisy population. Conversely, medication ordering errors that are "easily explained" (e.g., a student order, redundant order-set orders) can be filtered out, and more complex errors investigated in greater detail for safety and quality improvement purposes.

Conclusions and limitations

Our study has several limitations. The data was from a single academic medical center. The content findings are likely of limited generalizability. However, the methodology we have developed can be easily applied to similar datasets to identify contributors to medication ordering errors; a key advantage of ML approaches is that they are designed to be theory agnostic and account for tuning and discovery using split-sample methods. As described in our previous studies, not all voided orders were true errors. Based on chart review of voided orders in a previous study, we found that voided orders had a 60%-80% positive predictive value of being a true error [22, 34]. Approximately 22% of the self-intercepted voided orders reached the patient [34]. As voiding is optional for clinicians, certain types of orders may have been preferentially voided more frequently than others. For instance, clinicians may have preferentially voided duplicate student orders over physician orders, thus biasing the feature importance metrics. We would also like to highlight that the performance of these models is below what is needed for directly incorporating into clinical practice. For example, the false-positive rate of 90% to capture 20% of voided orders would create an unacceptable alarm burden (see Fig 1 right panel). Our dataset does not allow direct review of false-negative orders (unmarked errors); doing so would be impossible given the scale of orders that would be required. Our goal was not to replicate the domain knowledge dependent clinical decision support integrated into pharmacy systems, but to consider medication order-related factors that contribute to medication ordering errors. As such, our models did not consider details regarding the patient medical history, laboratory values, or other drug orders, which are central to most heuristic based CDS. Future studies can incorporate additional features of the

medication, patient, and prescriber to improve performance. As this study was based on historical data with different configurations and order sentences, we could not verify some of the settings of the order entry system. Finally, the considered models have very different degrees of interpretability in the discussed potential applications. Logistic regression and decision trees can be directly "read off" to identify the driving factors. Although many promising efforts [35, 36] are increasing the interpretability of other classifiers, these are inherently much more complex and vulnerable to misinterpretation. Future applications will have to balance their need for interpretability and the benefits of complexity.

Medication voiding offers a promising approach for detecting, tracking and organizing medication errors. Our findings, based on applying ML models to voided orders, highlight the potential for identifying common failures of CPOE use and finding orders likely to be errors based on contextual factors. This opens new opportunities for supplementation of clinical decision support development, real-time error surveillance of efficacy of innovations in CPOE, and pragmatic patient safety efforts.

Supporting information

S1 Checklist. TRIPOD checklist & calibration curves.
(DOCX)

Author Contributions

Conceptualization: Joanna Abraham, William Galanter, Thomas Kannampallil.

Data curation: Joanna Abraham, Zhicheng Cui, Thomas Kannampallil.

Formal analysis: Christopher Ryan King, Bradley A. Fritz, Zhicheng Cui.

Funding acquisition: Joanna Abraham.

Investigation: Joanna Abraham, Thomas Kannampallil.

Methodology: William Galanter.

Project administration: Thomas Kannampallil.

Supervision: Joanna Abraham, Yixin Chen.

Writing – original draft: Christopher Ryan King, Joanna Abraham, Zhicheng Cui, Thomas Kannampallil.

Writing – review & editing: Christopher Ryan King, Joanna Abraham, Bradley A. Fritz, William Galanter, Yixin Chen, Thomas Kannampallil.

References

1. Ammenwerth E, Schnell-Inderst P, Machan C, Siebert U. The effect of electronic prescribing on medication errors and adverse drug events: a systematic review. JAMIA. 2008; 15(5):585–600. https://doi.org/10.1197/jamia.M2667 PMID: 18579832
2. Bates DW, Leape LL, Cullen DJ, Laird N, Petersen LA, Teich JM, et al. Effect of computerized physician order entry and a team intervention on prevention of serious medication errors. Jama. 1998; 280 (15):1311–6. Epub 1998/10/30. https://doi.org/10.1001/jama.280.15.1311 PMID: 9794308.
3. Bates DW, Teich JM, Lee J, Seger D, Kuperman GJ, Ma'Luf N, et al. The impact of computerized physician order entry on medication error prevention. JAMIA. 1999; 6(4):313–21. https://doi.org/10.1136/jamia.1999.00660313 PMID: 10428004
4. Kuperman GJ, Gibson RF. Computer physician order entry: benefits, costs, and issues. Annals of internal medicine. 2003; 139(1):31. https://doi.org/10.7326/0003-4819-139-1-200307010-00010 PMID: 12834316

5. Prgomet M, Li L, Niazkhani Z, Georgiou A, Westbrook JI. Impact of commercial computerized provider order entry (CPOE) and clinical decision support systems (CDSSs) on medication errors, length of stay, and mortality in intensive care units: a systematic review and meta-analysis. Journal of the American Medical Informatics Association. 2016; 24(2):413–22.
6. van Rosse F, Maat B, Rademaker CM, van Vught AJ, Egberts AC, Bollen CW. The effect of computerized physician order entry on medication prescription errors and clinical outcome in pediatric and intensive care: a systematic review. Pediatrics. 2009; 123(4):1184–90. https://doi.org/10.1542/peds.2008-1494 PMID: 19336379
7. Potts AL, Barr FE, Gregory DF, Wright L, Patel NR. Computerized physician order entry and medication errors in a pediatric critical care unit. Pediatrics. 2004; 113(1):59–63.
8. Cordero L, Kuehn L, Kumar RR, Mekhjian HS. Impact of computerized physician order entry on clinical practice in a newborn intensive care unit. Journal of Perinatology. 2004; 24(2):88. https://doi.org/10.1038/sj.jp.7211000 PMID: 14872207
9. Bates DW. Using information technology to reduce rates of medication errors in hospitals. BMJ. 2000; 320(7237):788–91. https://doi.org/10.1136/bmj.320.7237.788 PMID: 10720369
10. Wetterneck TB, Walker JM, Blosky MA, Cartmill RS, Hoonakker P, Johnson MA, et al. Factors contributing to an increase in duplicate medication order errors after CPOE implementation. JAMIA. 2011; 18 (6):774–82. https://doi.org/10.1136/amiajnl-2011-000255 PMID: 21803925
11. Nuckols TK, Smith-Spangler C, Morton SC, Asch SM, Patel VM, Anderson LJ, et al. The effectiveness of computerized order entry at reducing preventable adverse drug events and medication errors in hospital settings: a systematic review and meta-analysis. Systematic Reviews. 2014; 3(1):56. https://doi.org/10.1186/2046-4053-3-56 PMID: 24894078
12. Shawahna R, Rahman NU, Ahmad M, Debray M, Yliperttula M, Decleves X. Electronic prescribing reduces prescribing error in public hospitals. Journal of Clinical Nursing. 2011; 20(21–22):3233–45. https://doi.org/10.1111/j.1365-2702.2011.03714.x PMID: 21627699
13. van Doormaal JE, van den Bemt PM, Zaal RJ, Egberts AC, Lenderink BW, Kosterink JG, et al. The Influence that Electronic Prescribing Has on Medication Errors and Preventable Adverse Drug Events: an Interrupted Time-series Study. Journal of the American Medical Informatics Association. 2009; 16 (6):816–25. https://doi.org/10.1197/jamia.M3099 PMID: 19717798
14. Abraham J, Kannampallil TG, Jarman A, Sharma S, Rash C, Schiff G, et al. Reasons for computerised provider order entry (CPOE)-based inpatient medication ordering errors: an observational study of voided orders. BMJ Quality & Safety. 2018; 27(4):299–307. https://doi.org/10.1136/bmjqs-2017-006606 PMID: 28698381
15. Korb-Savoldelli V, Boussadi A, Durieux P, Sabatier B. Prevalence of computerized physician order entry systems–related medication prescription errors: A systematic review. International Journal of Medical Informatics. 2018; 111:112–22. https://doi.org/10.1016/j.ijmedinf.2017.12.022 PMID: 29425622
16. Reckmann MH, Westbrook JI, Koh Y, Lo C, Day RO. Does computerized provider order entry reduce prescribing errors for hospital inpatients? A systematic review. JAMIA. 2009; 16(5):613–23. https://doi.org/10.1197/jamia.M3050 PMID: 19567798
17. Westbrook JI, Reckmann M, Li L, Runciman WB, Burke R, Lo C, et al. Effects of two commercial electronic prescribing systems on prescribing error rates in hospital in-patients: a before and after study. PLoS Medicine. 2012; 9(1):e1001164. https://doi.org/10.1371/journal.pmed.1001164 PMID: 22303286
18. Ghaleb MA, Barber N, Franklin BD, Yeung VW, Khaki ZF, Wong IC. Systematic Review of Medication Errors in Pediatric Patients. Annals of Pharmacotherapy. 2006; 40(10):1766–76. https://doi.org/10.1345/aph.1G717 PMID: 16985096.
19. Vrbnjak D, Denieffe S, O'Gorman C, Pajnkihar M. Barriers to reporting medication errors and near misses among nurses: A systematic review. International journal of nursing studies. 2016; 63:162–78. https://doi.org/10.1016/j.ijnurstu.2016.08.019 PMID: 27637011
20. Schiff G, Amato M, Eguale T, Boehne J, Wright A, Koppel R, et al. Computerised physician order entry-related medication errors: analysis of reported errors and vulnerability testing of current systems. BMJ Qual Saf. 2015; 24(4):264–71. https://doi.org/10.1136/bmjqs-2014-003555 PMID: 25595599
21. Hickman T-TT, Quist AJL, Salazar A, Amato MG, Wright A, Volk LA, et al. Outpatient CPOE orders discontinued due to 'erroneous entry': prospective survey of prescribers' explanations for errors. BMJ Qual Saf. 2018; 27(4):293–8. https://doi.org/10.1136/bmjqs-2017-006597 PMID: 28754812
22. Kannampallil TG, Abraham J, Solotskoya A, George S, Lambert BL, Schiff G, et al. Learning from Errors: Analysis of Medication Order Voiding in CPOE Systems Journal of the American Medical Informatics Association. 2017; 24(4):762–8. https://doi.org/10.1093/jamia/ocw187 PMID: 28339698
23. Liu Y, Wang Y, Wang S, Liang T, Zhao Q, Tang Z, et al., editors. CBNet: A Novel Composite Backbone Network Architecture for Object Detection. arXiv preprint arXiv:190903625; 2019.

24. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint 2014.

25. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning: ACM; 2006. p. 233–40.

26. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002; 16:321–57.

27. Qin G, Hotilovac L. Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test. Stat Meth Med Res. 2008; 17:207–21.

28. MatlabAUC. https://github.com/brian-lau/MatlabAUC.

29. Liao X, Meyer MC. cgam: An R Package for the Constrained Generalized Additive Model. 2019. 2019; 89(5):24. Epub 2019-05-09. https://doi.org/10.18637/jss.v089.i05

30. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for simplicity: The all convolutional net. arXiv preprint arXiv:141268062014.

31. Lundberg SM, Lee S-I, editors. A unified approach to interpreting model predictions. Advances in neural information processing systems; 2017.

32. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv:150302531. 2015.

33. Lundberg SM, Erion GG, Lee S-I. Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:180203888. 2018.

34. Abraham J, Galanter WL, Touchette D, Xia Y, Holzer KJ, Leung V, et al. Risk factors associated with medication ordering errors. J Am Med Inform Assoc. 2021; 28(1):86–94. Epub 2020/11/23. https://doi.org/10.1093/jamia/ocaa264 PMID: 33221852.

35. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. Nat Mach Intell. 2020; 2(1):56–67. Epub 2020/07/02. https://doi.org/10.1038/s42256-019-0138-9 PMID: 32607472.

36. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng. 2018; 2(10):749–60. Epub 2019/04/20. https://doi.org/10.1038/s41551-018-0304-0 PMID: 31001455.